DOCUMENT RESUME

ED 427 044 TM 029 422

AUTHOR Barnette, J. Jackson; McLean, James E.
TITLE Protected versus Unprotected Multiple Comparison Procedures.
PUB DATE 1998-11-05
NOTE 15p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (27th, New Orleans, LA, November 4-6, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; Monte Carlo Methods; *Research Methodology; Sample Size; Simulation
IDENTIFIERS *Bonferroni Procedure; Dunn Sidak Procedure; F Test; Holm Sequentially Rejective Procedure; Tukey Statistic; Type I Errors

ABSTRACT

Conventional wisdom suggests the omnibus F-test needs to be significant before conducting post-hoc pairwise multiple comparisons. However, there is little empirical evidence supporting this practice. Protected tests are conducted only after a significant omnibus F-test while unprotected tests are conducted without regard to the significance of the omnibus F-test. Monte Carlo methods were used to generate replications expected to provide 0.95 confidence intervals of +/- 0.001 around the nominal alphas of 0.10, 0.05, and 0.01 for 42 combinations of "n" (5, 10, 15, 20, 30, 60, and 100) and numbers of groups (3, 4, 5, 6, 8, and 10). Unprotected and protected tests were conducted using the Dunn-Bonferroni, Dunn-Sidak, Holm, and Tukey's Honestly Significant Differences (HSD) procedures. Means and standard deviations of observed per-experiment Type I errors rates and percentages of observed per-experiment Type I error falling below, within, and above the 0.95 confidence intervals were determined for total number of Type I errors. Differences in observed Type I errors for sample size and number of groups was minimal. However, there were differences in Type I error control among the four multiple comparison procedures and when the tests were conducted as protected or unprotected. The Dunn-Bonferroni had the best control of Type I error as an unprotected test with 96.0% of the observed Type I errors falling within the 0.95 confidence interval while 87.3% of the observed Type I errors fell below the 0.95 confidence interval when used as a protected test, thus being very conservative. As unprotected tests, the Dunn-Sidak and Holm tended to be liberal, but were conservative as protected tests. The HSD was liberal in both situations, but much more so as an unprotected test. These results, combined with the ease of using the Dunn-Bonferroni, suggest this method may provide the most accurate and easiest control of per-experiment Type I error when used in an unprotected mode. (Contains 4 tables and 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 427 044

Protected versus Unprotected Multiple Comparison Procedures

J. Jackson Barnette
University of Iowa

and

James E. McLean
University of Alabama at Birmingham

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Jackson Barnette

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Address correspondence to:

Dr. Jack Barnette
College of Medicine
1-204 Medical Education Building
University of Iowa
Iowa City, IA 52242-1000
(319) 335 8905
jack-barnette@uiowa.edu

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029422

A paper presented at the 1998 Annual meeting of the
Mid-South Educational Research Association, New Orleans, LA

November 5, 1998

Rev. 12 98

BEST COPY AVAILABLE

Abstract

Conventional wisdom suggests the omnibus F-test needs to be significant before conducting post-hoc pairwise multiple comparisons. However, there is little empirical evidence supporting this practice. Protected tests are conducted only after a significant omnibus F-test while unprotected tests are conducted without regard to the significance of the omnibus F-test.

Monte Carlo methods were used to generate replications expected to provide .95 confidence intervals of +/- .001 around the nominal alphas of .10, .05, and .01 for 42 combinations of *n* (5, 10, 15, 20, 30, 60, and 100) and numbers of groups (3, 4, 5, 6, 8, and 10). Unprotected and protected tests were conducted using the Dunn-Bonferroni, Dunn-Sidak, Holm, and Tukey's HSD procedures. Means and standard deviations of observed per-experiment Type I error-rates and percentages of observed per-experiment Type I errors falling below, within, and above the .95 CI's were determined for total number of Type I errors.

Differences in observed Type I errors for sample size and number of groups was minimal. However, there were differences in Type I error control among the four multiple comparison procedures and when the tests were conducted as protected or unprotected. The Dunn-Bonferroni had the best control of Type I error as an unprotected test with 96.0% of the observed Type I errors falling within the .95 CI while 87.3% of the observed Type I errors fell below the .95 CI when used as a protected test, thus being very conservative. As unprotected tests, the Dunn-Sidak and Holm tended to be liberal, but were conservative as protected tests. HSD was liberal in both situations, but much more so as an unprotected test. These results, combined with the ease of using the Dunn-Bonferroni, suggest this method may provide the most accurate and easiest control of per-experiment Type I error when used in an unprotected mode.

PROTECTED VERSUS UNPROTECTED MULTIPLE COMPARISON PROCEDURES

Conventional wisdom suggests the omnibus F-test needs to be significant before conducting post-hoc pairwise multiple comparisons in order to control Type I error rates. Many current texts support the practice for all post-hoc multiple comparison procedures. For example, Kirk (1982), indicates that "an overall F test is often merely the first step in analyzing a set of data" (p. 90). The practice, evidently, grew from Fisher's 1935 publication of the least significant difference (LSD) test where the omnibus F-test was the first step. The practice of performing multiple comparison tests only after a significant omnibus F-test has become known as "protected tests", while unprotected tests are conducted without regard to the significance of the omnibus F-test. While this has become common practice, particularly for post-hoc tests, there is little empirical evidence supporting it. The purpose of this paper is to examine this practice's impact on Type I errors, using Monte Carlo procedures with some of the more popular post-hoc multiple comparison procedures.

Background

Most studies of Type I error rates for identifying significant pairwise mean differences have been based on what is referred to as experimentwise or familywise error control philosophies. Experimentwise (EW) Type I error relates to finding at least one significant difference by chance for the specified alpha level. In these cases, the only difference of concern is the largest mean difference. Experimentwise Type I error control ignores the possibility of multiple Type I errors in the same experiment. The possibility of additional pairwise mean differences are not considered. Type I error control is such that not all possible Type I errors are evaluated. In these cases, many procedures such as Tukey's HSD are considered to have conservative Type I error control since the actual probabilities of finding at least one Type I error are lower than the nominal alpha level.

Per-experiment (PE) Type I error control considers all the possible Type I errors that can occur in a given experiment. Thus, more than one Type I error per experiment is possible and reasonably likely to occur if there is an experimentwise Type I error on the highest mean difference. Klockars and Hancock (1994) pointed out the importance and risks associated with this distinction. They found, using a Monte Carlo simulation, that there was a difference of .0132 in the per-experiment and experimentwise Type I error rates for Tukey's HSD when alpha was set at .05. Thus, when one has exact control of Type I error in the experimentwise situation, the per-experiment Type I error probability is higher. While most Type I error research is based on an experimentwise mode, the per-experiment Type I error is more consistent with the reality of pairwise hypothesis testing. It is not only the largest mean difference subjected to error control but all the pairwise differences. Thus, we favor a per-experiment mode of Type I error control; that is what is used in this research.

Four multiple comparison procedures were selected for this research: Dunn-Bonferroni, Dunn-Sidak, Holm's sequentially rejective, and Tukey's HSD. Based on a review of current literature and commonly used statistical texts, we have concluded that these are among the most frequently used pairwise procedures and represent a variety of

approaches to control for Type I error. Since the names of these procedures tend to vary slightly in texts, statistical software, and in the literature, each is described briefly below:

Dunn-Bonferroni Procedure. The Dunn-Bonferroni procedure uses the Bonferroni inequality ($\alpha_{PE} \leq \sum \alpha_{PC}$) as authority to divide equally the total a priori error among the number of tests to be completed, often following the application of the Fisher LSD procedure. The LSD procedure is equivalent to conducting all pairwise comparisons using independent t-tests with the MS_{error} as the common pooled variance estimate (Kirk, 1982). An example of the application of the Dunn-Bonferroni would be identifying the a priori α as .05 where tests are required to compare means of five groups using 10 comparisons, running each individual test at the .05/10= .005 level (Hayes, 1988).

Dunn-Sidak Procedure. Sidak's modification of the Dunn-Bonferroni Procedure substituted the multiplicative computation of the exact error-rate, $\alpha_{PE} = 1 - (1 - \alpha_{PC})^c$ where c is the number of comparisons for the Bonferroni Inequality ($\alpha_{PE} \leq \sum \alpha_{PC}$), otherwise following the same procedures (Kirk, 1982).

Holm's Sequentially Rejective Procedure. This procedure was proposed by Holm in 1979 and is also referred to as the Sequentially Rejective Bonferroni Procedure. Assuming a maximum of *c* comparisons to be performed, the first null hypothesis is tested at the α/c level. If the test is significant, the second null hypothesis is tested at the $\alpha/(c - 1)$ level. If this is significant, the testing continues in a similar manner until all *c* tests have been completed or until a nonsignificant test is run. The testing stops when the first nonsignificant test is encountered (Hancock & Klockars, 1996).

Tukey's Honestly Significant Difference Procedure (HSD). This procedure was presented originally in a non-published paper by Tukey in 1953. Its popularity has grown to the point where it is, possibly, the most widely used multiple comparison procedure. The HSD is based on the Studentized Range Statistic originally derived by Gosset (a.k.a., Student) (1907-1938). This statistic, unlike the t-statistic, takes into account the number of means being compared, adjusting for the total number of tests to make all pairwise comparisons (Kennedy & Bush, 1985). Purportedly, the HSD gives a per-experiment error rate.

Many common statistical texts either recommend or imply the use of a protected test for all post-hoc multiple comparison procedures (e.g., Hayes, 1988; Kennedy & Bush, 1985; Kirk, 1982; Maxwell & Delaney, 1990). While these texts provide a logical basis for this, and excellent reviews of multiple comparison procedures are available (e.g., Hancock & Klockars, 1996; Toothaker, 1993), little empirical evidence is presented. No study could be found that derived, either analytically or empirically, a justification for this practice.

Methodology

Monte Carlo methods were used to generate the data for this research. All data comprising the groups whose means were compared were generated from a random normal deviate routine, which was incorporated into a larger compiled QBASIC program

that conducted all needed computations. The program was written by the senior author. All sampling and computation, conducted with double-precision, routines were verified using SAS® programs. The program was run on a Dell Pentium II, 266 MHZ personal computer. A more detailed description of this process can be found in Barnette and McLean (1997b, November); an illustration of how they were applied to the testing of another multiple comparison procedure can be found in Barnette and McLean (1997a, November).

Several sample size and number of groups arrangements were selected to give a range of low, moderate, and large case situations. The number of groups was: 3, 4, 5, 6, 8, and 10 and the sample sizes for each group were: 5, 10, 15, 20, 30, 60, and 100, which when crossed gave 42 experimental conditions. This was replicated for three nominal alphas of .10, .05, and .01. The approach used was to determine what number of replications would be needed to provide an expected .95 confidence interval of +/- .001 around the nominal alpha. This is an approach to examination of how well observed Type I error proportions are reasonable estimates of a standard nominal alpha. In other words, if alpha is the standard, what proportion of the estimates of actual Type I error proportions can be considered accurate, as evidenced by them being within the expected .95 confidence interval around nominal alpha?

This was based on the assumption that errors would be normally distributed around the binomial proportion represented by nominal alpha. Thus, when alpha was .10, 345742 replications were needed to have a .95 confidence interval of +/- .001 or between .099 and .101. When alpha was .05, 182475 replications were needed to have a .95 confidence interval of +/- .001 or between .049 and .051 and when alpha was .01, 38032 replications were needed to have a .95 confidence interval of +/- .001 or between .009 and .011. Observed Type I error proportions falling into the respective .95 confidence intervals are considered to be reliable estimates of the expected Type I error rate. Observed Type I error proportions falling below the .95 CI are considered to be conservative; observed Type I error proportions falling above the .95 CI are considered to be liberal.

Within each nominal alpha/sample size/number of groups configuration, the number of ANOVA replications were generated. Each replication involved drawing of elements of the sample from a distribution of normal deviates, computation of sample means, and the omnibus F test. The four multiple comparison procedures were conducted without regard to the results of the omnibus F test, the unprotected mode; then they were conducted again using the same data if the omnibus F statistic was significant, the protected mode. Mean differences were examined for all differences from the K number of steps down to 2 steps between ordered means. This approach permitted determination of per-experiment Type I errors or the total number of Type I errors observed, regardless of where they are in the stepwise structure. Summary statistics were computed for unprotected and protected conditions for each alpha level including: the mean proportion of per-experiment Type I errors, standard deviation of the proportions of per-experiment Type I errors, minimum proportion, maximum proportion, and the percentage of proportions falling in the three regions associated with the .95 confidence interval.

Results

First, the results are presented separately for alpha levels of .10, .05, and .01. Then, the results are summarized in terms of trends and comparisons to the preset nominal alpha levels.

Per-Experiment Error, Alpha of .10

Table 1 presents the results when alpha is set at .10. When 345742 samples are used, the expected .95 confidence interval around .10 is +/- .001 or from .099 to .101. Observed probabilities in that range could be considered to be accurate estimates of alpha of .10. Values falling below the confidence interval could be considered conservative, the further below .10, the more conservative. Values above the confidence interval could be considered to be liberal, the further above .10, the more liberal. These results represent the summary for the 42 "number of groups" and "group size" configurations.

Results for the Dunn-Bonferroni indicate a mean probability of .1002 when conducted as an unprotected test and .0947 when conducted as a protected test. There was lower variance for the unprotected condition compared with the protected condition. In the unprotected condition 92.9% of the probabilities were within the expected .95 confidence interval, while 81.0% of the probabilities were below the expected .95 confidence interval when conducted as a protected test. Thus, the Dunn-Bonferroni procedure demonstrated accurate control of per-experiment Type I error as an unprotected test, while as a protected test it was conservative.

The mean probability for the Dunn-Sidak as an unprotected test was .1049 with 100% of the probabilities above the expected .95 confidence interval. This indicates the Dunn-Sidak is liberal as an unprotected test. As a protected test, the mean probability was .0985 with higher variance of observed probabilities. For the protected test, 47.6% of the observed probabilities were below the .95 confidence interval, 19.0% were within the confidence interval, and 33.3% were above the confidence interval. Clearly there was more variance in the probabilities when used as a protected test compared with the unprotected test. Examination of the protected probabilities for the group number/group size configurations revealed that the test became more conservative as the number of groups increased. Thus, as an unprotected test across all group numbers and group sizes and as a protected test for lower numbers of groups (fewer than five), the Dunn-Sidak is liberal, but as a protected test with more than five groups it tends to be conservative.

Holm's sequentially rejective procedure, when used as an unprotected test, had a mean of .1059 and as a protected test had a mean of .1005. The protected test had more variance than the unprotected test. As an unprotected test, the Holm had 95.2% of the probabilities above the confidence interval and only 4.8% within the confidence interval, indicating a liberal test. When used as a protected test, 47.6% had probabilities below the confidence interval, 2.4% were within the confidence interval, and 50% were above the confidence interval. Examination of the means across the number of groups and group size configurations indicated that as group size increases, the Holm procedure becomes more conservative in control of per-experiment Type I errors. Thus, when alpha is set at .10, Holm's procedure is liberal when conducted as an unprotected test or when used as a

protected test with five or fewer groups; and it is conservative when used as a protected test with more than five groups.

Tukey's HSD had a mean probability of .1464 as an unprotected test and .1279 as a protected test. The variance was higher when used as an unprotected test compared with its use as a protected test. Of the four methods tested, only the HSD had a higher variance as an unprotected test. In both testing configurations, 100% of the probabilities were above the confidence interval. Thus, as a method of control of per-experiment Type I error when alpha is .10, the HSD is liberal, more liberal as an unprotected test.

In summary, when alpha was .10 the procedure with the best control of per-experiment Type I error was the Dunn-Bonferroni conducted as an unprotected test. However, the Dunn-Bonferroni was conservative as a protected test. The Dunn-Sidak, Holm, and HSD procedures were liberal when conducted as unprotected tests. The Dunn-Sidak and Holm procedures were liberal as protected tests with smaller numbers of groups, but tended to become conservative as larger number of groups were used; the HSD was liberal as a protected test in all number of groups and group size conditions.

Per-Experiment Error, Alpha of .05

Table 2 presents the results when alpha is set at .05. When 182475 samples are used, the expected .95 confidence interval around .05 is +/- .001 or from .049 to .051. Observed probabilities in that range could be considered to be accurate estimates of alpha of .05. Values falling below the confidence interval could be considered conservative, the further below .05, the more conservative. Values above the confidence interval could be considered to be liberal, the further above .05, the more liberal. These results represent the summary for the 42 "number of groups" and "group size" configurations.

Results for the Dunn-Bonferroni indicate a mean probability of .0500 when conducted as an unprotected test and .0448 when conducted as a protected test. There was lower variance for the unprotected condition compared with the protected condition. In the unprotected condition 95.2% of the probabilities were within the expected .95 confidence interval, while 92.9% of the probabilities were below the expected .95 confidence interval when conducted as a protected test. Thus, the Dunn-Bonferroni procedure demonstrated accurate control of per-experiment Type I error as an unprotected test, while as a protected test it was conservative.

The mean probability for the Dunn-Sidak as an unprotected test was .0511 with 50% of the probabilities above the expected .95 confidence interval and 50% above the confidence interval. This indicates the Dunn-Sidak tended to be accurate to liberal as an unprotected test, becoming more liberal as the number of groups and sample size increased. As a protected test, the mean probability was .0456 with higher variance of observed probabilities. For the protected test, 83.3% of the observed probabilities were below the .95 confidence interval and 16.7% were within the confidence interval. Clearly there was more variance in the probabilities when used as a protected test compared with the unprotected test. Examination of the protected test probabilities for the group number/group size configurations revealed that the test became more conservative as the number of groups increased.

Holm's sequentially rejective procedure, when used as an unprotected test, had a mean of .0521 and, as a protected test, had a mean of .0470. The protected test had more variance than the unprotected test. As an unprotected test, the Holm had 66.7% of the probabilities above the confidence interval and 33.3% within the confidence interval, indicating a liberal test. This test tended to be more accurate when larger number of groups were used. When used as a protected test, 61.9% had probabilities below the confidence interval, 19.0% were within the confidence interval, and 19.0% were above the confidence interval. Examination of the means across the number of groups and group size configurations indicated that when there were three groups, the test tended to be liberal. When there were four groups the test was accurate, but when there were more than four groups, the Holm procedure became more conservative in control of per-experiment Type I errors. Thus, when alpha is set at .05, Holm's procedure tends to be liberal when conducted as an unprotected test, but tends to be conservative when used as a protected test.

Tukey's HSD had a mean probability of .0667 as an unprotected test and .0553 as a protected test. The variance was higher when used as an unprotected test compared with its use as a protected test. Of the four methods tested, only the HSD had a higher variance as an unprotected test. As an unprotected test, 100% of the probabilities were above the .95 confidence interval; when tested as a protected test, 97.6% were above the confidence interval. Thus, as a method of control of per-experiment Type I error when alpha is .05, the HSD is liberal, more liberal as an unprotected test.

In summary, when alpha was .05 the procedure with the best control of per-experiment Type I error was the Dunn-Bonferroni conducted as an unprotected test. However, the Dunn-Bonferroni was conservative as a protected test. The Dunn-Sidak, Holm, and HSD procedures were liberal conducted as unprotected tests, with HSD being the most liberal. The Dunn-Sidak and Holm procedures tended to be conservative as protected tests, the Dunn-Sidak being more conservative. The HSD was liberal as a protected test in all "number of groups" and "group size" conditions.

Per-Experiment Error, Alpha of .01

Table 3 presents the results when alpha is set at .01. When 38032 samples are used, the expected .95 confidence interval around .01 is +/- .001 or from .009 to .011. Observed probabilities in that range could be considered to be accurate estimates of alpha of .01. Values falling below the confidence interval could be considered conservative, the further below .01, the more conservative. Values above the confidence interval could be considered to be liberal, the further above .01, the more liberal. These results represent the summary for the 42 "number of groups" and "group size" configurations.

Results for the Dunn-Bonferroni indicate a mean probability of .0100 when conducted as an unprotected test and .0079 when conducted as a protected test. There was lower variance for the unprotected condition compared with the protected condition. In the unprotected condition 100.0% of the probabilities were within the expected .95 confidence interval, while 88.1% of the probabilities were below the expected .95 confidence interval when conducted as a protected test. Thus, the Dunn-Bonferroni

procedure demonstrated accurate control of per-experiment Type I error as an unprotected test, while as a protected test it was conservative.

The mean probability for the Dunn-Sidak as an unprotected test was .0101 with 95.2% of the probabilities within the expected .95 confidence interval. This indicates the Dunn-Sidak is relatively accurate as an unprotected test. As a protected test, the mean probability was .0079 with higher variance of observed probabilities. For the protected test, 85.7% of the observed probabilities were below the .95 confidence interval and 14.3% were within the confidence interval. Thus, as an unprotected test across all group numbers and group sizes the Dunn-Sidak is accurate; but as a protected test it tends to be conservative in control of per-experiment Type I error.

Holm's sequentially rejective procedure, when used as an unprotected test, had a mean of .0103 and as a protected test had a mean of .0082. The protected test had more variance than the unprotected test. As an unprotected test, the Holm had 95.2% of the probabilities within the confidence interval and only 4.8% above the confidence interval, indicating a relatively accurate test. When used as a protected test, 69.0% had probabilities below the confidence interval and 31.0% within the confidence interval. Examination of the means across the number of groups and group size configurations indicated that as the number of groups increases, the Holm procedure, when used as a protected test, becomes more conservative in control of per-experiment Type I errors. Thus, when alpha is set at .01, Holm's procedure is accurate when conducted as an unprotected test but tends to become conservative when used as a protected test with larger numbers of groups.

Tukey's HSD had a mean probability of .0118 as an unprotected test and .0088 as a protected test. In this case, contrary to the cases when alpha was .10 or .05, the variance was higher when used as a protected test compared with its use as an unprotected test. When conducted as an unprotected test, 85.7% of the probabilities were above the confidence interval and 14.3% were within the confidence interval. The probabilities within the confidence interval were when the number of groups was three. Contrary to the situations where alpha was .10 or .05, there were 57.1% of the probabilities below the confidence interval while 42.9% were within the confidence interval. In situations with fewer groups with lower group sizes, HSD was more accurate, but became more conservative when more groups with larger group sizes were used Thus, as a method of control of per-experiment Type I error when alpha is .01, the HSD is liberal when conducted as an unprotected test, but tends to be conservative when used as a protected test.

In summary, when alpha was .01 the procedure with the best control of per-experiment Type I error was the Dunn-Bonferroni conducted as an unprotected test. However, the Dunn-Bonferroni was conservative as a protected test. The Dunn-Sidak and Holm procedures were relatively accurate when conducted as unprotected tests, but were conservative when used as protected tests. HSD was liberal when conducted as an unprotected test but tended to be accurate or conservative when conducted as a protected test.

Summary Across Alpha Levels

Table 4 summarizes the percentages of probabilities falling below, within, and above the .95 confidence intervals across the three levels of alpha. It is clear that overall there is only one, with 96.0% of the probabilities within the confidence interval, consistently accurate method that controls for per-experiment Type I error: the Dunn-Bonferroni when conducted as an unprotected test. Other methods with accurate or slightly liberal results were the Dunn-Sidak and Holm conducted as unprotected tests. Tukey's HSD was very liberal when used as an unprotected test and somewhat less liberal, although quite varied, when used as a protected test. The Dunn-Bonferroni or Dunn-Sidak methods, when used as protected tests, were very conservative. Holm's procedure demonstrated high variability as a protected test with 59.5% of the probabilities being below the confidence intervals, 17.5% within the confidence intervals, and 23.0% being above the confidence intervals.

It was unexpected to see that for some of the tests, the percentages falling into the areas relative to the confidence interval changed as alpha changed. For the Dunn-Bonferroni, the relative percentages remained about the same for all three alpha levels conducted as either unprotected or protected tests. As alpha decreased from .10 to .01, the trend was for the other three tests to become more conservative. For the Dunn-Sidak and Holm procedures when conducted as unprotected tests, the trend was for the tests to move from being very liberal to being relatively accurate. However, when conducted as protected tests, these two methods moved toward being even more conservative as alpha decreased. HSD when conducted as an unprotected test remained liberal in per-experiment Type I error control across all alpha levels. However, HSD moved from being very liberal at alpha .10 and .05 to being conservative at alpha of .01, when conducted as a protected test.

Conclusions

These results provide insights on two major controversies. One is the use of experimentwise vs. per-experiment Type I error control. We contend that per-experiment mode is closest to the realities of pairwise hypothesis testing, since more than just the largest pairwise difference is of interest and all pairwise comparisons are tested. The conventional wisdom, based on experimentwise Type I error control, is that the Dunn-Bonferroni is very conservative and that the HSD is conservative, but less so. The HSD is often recommended because it is conservative, yet provides reasonable power for finding significant differences; but this relates to experimentwise control and a protected test. Yet, arguments could be made that the HSD gets its power from a higher-than-nominal alpha level. In our research, when HSD is used as a protected test with alpha of .10 or .05, the actual per-experiment Type I error rates are .1279 and .0553 respectively. Thus, the operational alpha level is not the nominal level, but a higher level. If one truly is interested in maintaining an accurate level of control of Type I error, then methods which are shown to provide accurate actual controls should be used, and the power available can be determined by other comparison conditions: sample size, effect size, number of groups, and error variance.

The other controversy is the need for a significant omnibus F test as the gateway for conducting pairwise follow-ups. Is it not possible, as Hancock and Klockars (1996)

point out, that this requirement overprotects against finding pairwise differences? This research indicates that if you desire accurate, as advertised, control of per-experiment Type I error, there is one method that seems to provide that regardless of alpha level, number of groups, or number of subjects: the **Dunn-Bonferroni conducted as an unprotected test**. An unprotected Dunn-Bonferroni is more powerful than a protected Dunn-Bonferroni. Preliminary comparisons, using the data we have collected and used in this paper, indicate that a Dunn-Bonferroni conducted as an unprotected test loses little power compared with HSD used as a protected test. Another advantage to using an unprotected Dunn-Bonferroni is the ease of computation compared with any of the other methods. No tables or statistical analysis beyond the determination of pairwise t-test probabilities are needed.

We realize these findings are not consistent with common wisdom or with recommendations found in most statistics texts. However, we hope this research influences others to replicate our work, possibly using other methods. Only when we are willing to question our current practice are we able to improve on it.

References

Barnette, J. J., & McLean, J. E. (1997a, November). *A Comparison of Type I Error Rates of Alpha-Max with Established Multiple Comparison Procedures*. Paper presented at the annual meeting of the Mid-South Educational Research Association. Memphis, TN.

Barnette, J. J., & McLean, J. E. (1997a, November). *Using Monte Carlo methods for methodological research*. Training Session presented at the annual meeting of the Mid-South Educational Research Association. Memphis, TN.

Fisher, R. A. (1935, 1960). The design of experiments, 7th ed. London: Oliver & Boyd; New York: Hafner.

Gosset, W. S. (1907-1938) (1943). Student's collected papers. (E. S. Pearson & Wishart, J., editors). London: University Press, Biometrika Office.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971), *Review of Educational Research, 66*, 269-306.

Hayes, W. L. (1988). *Statistics* (4th ed). New York: Holt, Rinehart, and Winston, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Kennedy, J. J., & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America, Inc.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. (2nd ed). Belmont, CA: Brooks Cole.

Klockars, A. J. & Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement, 54* (2), 292-298.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing Company.

Toothaker, L.E. (1993). *Multiple comparison procedures*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-089. Newbury Park, CA: Sage.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Unpublished manuscipt, Princeton University.

Table 1. Summary Statistics and Percent of Observed Probabilities of Per-Experiment Type I Errors Below, Within, and Above CI_{95} (±.001) Across Nominal Alpha of .10 for Four Multiple Comparison Procedures Conducted as Unprotected and Protected Tests.

MCP	Conducted as:	Mean	SD	Min.	Max.	% below CI	% within CI	% above CI
Dunn-Bonferroni	Unprotected	.100160	.000661	.0997	.1020	0	92.9	7.1
	Protected	.094683	.004222	.0860	.1004	81.0	19.0	0
Dunn-Sidak	Unprotected	.104845	.000848	.1032	.1072	0	0	100.0
	Protected	.098462	.003845	.0897	.1039	47.6	19.0	33.3
Holm	Unprotected	.105864	.003385	.1007	.1127	0	4.8	95.2
	Protected	.100486	.007226	.0876	.1127	47.6	2.4	50.0
Tukey's HSD	Unprotected	.146410	.015567	.1206	.1790	0	0	100.0
	Protected	.127948	.009794	.1125	.1516	0	0	100.0

Table 2. Summary Statistics and Percent of Observed Probabilities of Per-Experiment Type I Errors Below, Within, and Above CI_{95} (±.001) Across Nominal Alpha of .05 for Four Multiple Comparison Procedures Conducted as Unprotected and Protected Tests.

MCP	Conducted as:	Mean	SD	Min.	Max.	% below CI	% within CI	% above CI
Dunn-Bonferroni	Unprotected	.049974	.000533	.0487	.0510	4.8	95.2	0
	Protected	.044829	.003152	.0385	.0501	92.9	7.1	0
Dunn-Sidak	Unprotected	.051105	.000520	.0500	.0524	0	50.0	50.0
	Protected	.045595	.003086	.0394	.0507	83.3	16.7	0
Holm	Unprotected	.052091	.001464	.0495	.0551	0	33.3	66.7
	Protected	.046962	.004329	.0391	.0551	61.9	19.0	19.0
Tukey's HSD	Unprotected	.066729	.005407	.0574	.0802	0	0	100.0
	Protected	.055310	.003240	.0504	.0640	0	2.4	97.6

Table 3. Summary Statistics and Percent of Observed Probabilities of Per-Experiment Type I Errors Below, Within, and Above CI_{95} (±.001) Across Nominal Alpha of .01 for Four Multiple Comparison Procedures Conducted as Unprotected and Protected Tests.

MCP	Conducted as:	Mean	SD	Min.	Max.	% below CI	% within CI	% above CI
Dunn Bonferroni	Unprotected	.010041	.000455	.0093	.0110	0	100.0	0
	Protected	.007912	.000999	.0062	.0096	88.1	11.9	0
Dunn-Sidak	Unprotected	.010081	.000470	.0093	.0111	0	95.2	4.8
	Protected	.007941	.000999	.0062	.0097	85.7	14.3	0
Holm	Unprotected	.010274	.000512	.0094	.0115	0	95.2	4.8
	Protected	.008157	.001163	.0062	.0102	69.0	31.0	0
Tukey's HSD	Unprotected	.011807	.000805	.0101	.0137	0	14.3	85.7
	Protected	.008772	.000977	.0069	.0105	57.1	42.9	0

Table 4. Percent of Observed Per-Experiment Type I Errors Below, Within, and Above CI_{95} (±.001) Across Nominal Alphas of .10, .05, and .01 for Four Multiple Comparison Procedures Conducted as Unprotected and Protected Tests

MCP	Conducted as:	% below CI	% within CI	% above CI
Dunn-Bonferroni	Unprotected	1.6	96.0	2.4
	Protected	87.3	12.7	0
Dunn-Sidak	Unprotected	0	48.4	51.6
	Protected	85.7	14.3	0
Holm	Unprotected	0	44.4	55.6
	Protected	59.5	17.5	23.0
Tukey's HSD	Unprotected	0	4.8	95.2
	Protected	19.0	15.1	65.9